Attachment 3:

Questions used to guide the development of criteria for each domain in experimental animal toxicology studies

Evaluation type	Domain– core question	Prompting questions	Basic considerations
Reporting Quality	**Reporting Quality –** Does the study report information for evaluating the design and conduct of the study for the endpoint(s)/outcome(s) of interest? *Notes:* *Reviewers should reach out to authors to obtain missing information when studies are considered key for hazard evaluation and/or dose-response.* *This domain is limited to reporting. Other aspects of the exposure methods, experimental design, and endpoint evaluation methods are evaluated using the domains related to risk of bias and study sensitivity.*	Does the study report the following? **Critical information** necessary to perform study evaluation: Species; test article name; levels and duration of exposure; route (e.g., oral; inhalation); qualitative or quantitative results for at least one endpoint of interest. **Important information** for evaluating the study methods: Test animal: strain, sex, source, and general husbandry procedures. Exposure methods: source, purity, method of administration. Experimental design: frequency of exposure, animal age and lifestage during exposure and at endpoint/outcome evaluation. Endpoint evaluation methods: assays or procedures used to measure the endpoints/outcomes of interest.	These considerations typically do not need to be refined by assessment teams, although in some instances the **important information** may be refined depending on the endpoints/outcomes of interest or the chemical under investigation. A judgment and rationale for this domain should be given for the study. Typically, these will not change regardless of the endpoints/outcomes investigated by the study. **In the rationale, reviewers should indicate whether the study adhered to GLP, OECD, or other testing guidelines.** *Good*: All critical and **important information** is reported or inferable for the endpoints/outcomes of interest. *Adequate*: All **critical information** is reported but some **important information** is missing. However, the missing information is not expected to significantly impact the study evaluation. *Deficient*: All **critical information** is reported but **important information** is missing that is expected to significantly reduce the ability to evaluate the study. *Critically Deficient*: Study report is missing any pieces of **critical information**. Studies that are Critically Deficient for reporting are Uninformative for the overall rating and not considered further for evidence synthesis and integration.

Questions used to guide the development of criteria for each domain in experimental animal toxicology studies

Evaluation type		Domain– core question	Prompting questions	Basic considerations
Risk of Bias	Selection and performance bias	**Allocation –** Were animals assigned to experimental groups using a method that minimizes selection bias?	For each study: Did each animal or litter have an equal chance of being assigned to any experimental group (i.e., random allocation)? Is the allocation method described? Aside from randomization, were any steps taken to balance variables across experimental groups during allocation?	These considerations typically do not need to be refined by assessment teams. A judgment and rationale for this domain should be given for each cohort or experiment in the study. *Good*: Experimental groups were randomized and any specific randomization procedure was described or inferable (e.g., computer-generated scheme). [Note that normalization is not the same as randomization (see response for 'Adequate').]. *Adequate*: Authors report that groups were randomized but do not describe the specific procedure used (e.g., "animals were randomized"). Alternatively, authors used a non-random method to control for important modifying factors across experimental groups (e.g., body weight normalization). *Not Reported* (interpreted as Deficient): No indication of randomization of groups or other methods (e.g., normalization) to control for important modifying factors across experimental groups. *Critically Deficient*: Bias in the animal allocations was reported or inferable.
		Observational bias/Blinding– Did the study implement measures to reduce observational bias?	For each endpoint/outcome or grouping of endpoints/outcomes in a study: Does the study report blinding or other methods/procedures for reducing observational bias? If not, did the study use a design or approach for which such procedures can be inferred? What is the expected impact of failure to implement (or report implementation) of these methods/procedures on results?	These considerations typically do not need to be refined by the assessment teams. [Note that it can be useful for teams to identify highly subjective measures of endpoints/outcomes where observational bias may strongly influence results prior to performing evaluations.] A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. *Good*: Measures to reduce observational bias were described (e.g., blinding to conceal treatment groups

Questions used to guide the development of criteria for each domain in experimental animal toxicology studies

Evaluation type		Domain– core question	Prompting questions	Basic considerations
				during endpoint evaluation; consensus-based evaluations of histopathology lesions).[a] *Adequate*: Methods for reducing observational bias (e.g., blinding) can be inferred or were reported but described incompletely. *Not Reported*: Measures to reduce observational bias were not described. (interpreted as Adequate) The potential concern for bias was mitigated based on use of automated/computer driven systems, standard laboratory kits, relatively simple, objective measures (e.g., body or tissue weight), or screening-level evaluations of histopathology. (interpreted as Deficient) The potential impact on the results is major (e.g., outcome measures are highly subjective). *Critically Deficient*: Strong evidence for observational bias that could have impacted results.
	Confounding/ variable control	**Confounding–** Are variables with the potential to confound or modify results controlled for and consistent across all experimental groups?	For each study: Are there differences across the treatment groups (e.g., co-exposures, vehicle, diet, palatability, husbandry, health status, etc.) that could bias the results? If differences are identified, to what extent are they expected to impact the results?	These considerations may need to be refined by assessment teams, as the specific variables of concern can vary by experiment or chemical. A judgment and rationale for this domain should be given for each cohort or experiment in the study, noting when the potential for confounding is restricted to specific endpoints/outcomes. *Good*: Outside of the exposure of interest, variables that are likely to confound or modify results appear to be controlled for and consistent across experimental groups. *Adequate*: Some concern that variables that were likely to confound or modify results were uncontrolled or

Questions used to guide the development of criteria for each domain in experimental animal toxicology studies

Evaluation type		Domain– core question	Prompting questions	Basic considerations
				inconsistent across groups, but are expected to have a minimal impact on the results. *Deficient*: Notable concern that potentially confounding variables were uncontrolled or inconsistent across groups, and are expected to substantially impact the results. *Critically deficient*: Confounding variables were presumed to be uncontrolled or inconsistent across groups, and are expected to be a primary driver of the results.
	Reporting and attrition bias	**Selective reporting and attrition–** Did the study report results for all prespecified outcomes and tested animals? *Note:* *This domain does* ***not*** *consider the appropriateness of the analysis/results presentation. This aspect of study quality is evaluated in another domain.*	For each study: *Selective reporting bias:* Are all results presented for endpoints/outcomes described in the methods (see note)? *Attrition bias:* Are all animals accounted for in the results? If there are discrepancies, do authors provide an explanation (e.g., death or unscheduled sacrifice during the study)? If unexplained results omissions and/or attrition are identified, what is the expected impact on the interpretation of the results?	These considerations typically do not need to be refined by assessment teams. A judgment and rationale for this domain should be given for each cohort or experiment in the study. *Good*: Quantitative or qualitative results were reported for all prespecified outcomes (explicitly stated or inferred), exposure groups and evaluation timepoints. Data not reported in the primary article is available from supplemental material. If results omissions or animal attrition are identified, the authors provide an explanation and these are not expected to impact the interpretation of the results. *Adequate*: Quantitative or qualitative results are reported for most prespecified outcomes (explicitly stated or inferred), exposure groups and evaluation timepoints. Omissions and/or attrition are not explained, but are not expected to significantly impact the interpretation of the results. *Deficient*: Quantitative or qualitative results are missing for many prespecified outcomes (explicitly stated or inferred), exposure groups and evaluation timepoints and/or high animal attrition; omissions

Questions used to guide the development of criteria for each domain in experimental animal toxicology studies

Evaluation type		Domain– core question	Prompting questions	Basic considerations
				and/or attrition are not explained and may significantly impact the interpretation of the results. *Critically Deficient*: Extensive results omission and/or animal attrition are identified and prevents comparisons of results across treatment groups.
Sensitivity	Exposure methods sensitivity	**Chemical administration and characterization–** Did the study adequately characterize exposure to the chemical of interest and the exposure administration methods? *Note:* *Consideration of the appropriateness of the route of exposure is not evaluated at the individual study level. Relevance and utility of the routes of exposure are considered in the PECO criteria for study inclusion and during evidence synthesis.*	For each study: Does the study report the source and purity and/or composition (e.g., identity and percent distribution of different isomers) of the chemical? If not, can the purity and/or composition be obtained from the supplier (e.g., as reported on the website). Was independent analytical verification of the test article purity and composition performed? Did the authors take steps to ensure the reported exposure levels were accurate? For inhalation studies: were target concentrations confirmed using reliable analytical measurements in chamber air? For oral studies: if necessary based on consideration of chemical-specific knowledge (e.g., instability in solution; volatility) and/or exposure design (e.g., the frequency and duration of exposure), were chemical concentrations in the dosing solutions or diet analytically confirmed? Are there concerns about the methods used to administer the chemical (e.g., inhalation chamber type, gavage volume, etc.)?	It is essential that these criteria are considered, and potentially refined, by assessment teams, as the specific variables of concern can vary by chemical. A judgment and rationale for this domain should be given for each cohort or experiment in the study. *Good*: Chemical administration and characterization is complete (i.e., source, purity, and analytical verification of the test article are provided). There are no concerns about the composition, stability, or purity of the administered chemical, or the specific methods of administration. For inhalation studies, chemical concentrations in the exposure chambers are verified using reliable analytical methods. *Adequate*: Some uncertainties in the chemical administration and characterization are identified but these are expected to have minimal impact on interpretation of the results (e.g., source and vendor-reported purity are presented, but not independently verified; purity of the test article is sub-optimal but not concerning; For inhalation studies, actual exposure concentrations are missing or verified with less reliable methods). *Deficient*: Uncertainties in the exposure characterization are identified and expected to substantially impact the results (e.g., source of the test article is not reported; levels of impurities are substantial or concerning; deficient administration methods, such as use of static inhalation chambers or a gavage

Questions used to guide the development of criteria for each domain in experimental animal toxicology studies

Evaluation type		Domain– core question	Prompting questions	Basic considerations
				volume considered too large for the species and/or lifestage at exposure). *Critically Deficient*: Uncertainties in the exposure characterization are identified and there is reasonable certainty that the results are largely attributable to factors other than exposure to the chemical of interest (e.g., identified impurities are expected to be a primary driver of the results).
		Exposure timing, frequency and duration– Was the was the timing, frequency, and duration of exposure sensitive for the endpoint(s)/outcome(s) of interest?	For each endpoint/outcome or grouping of endpoints/outcomes in a study: Does the exposure period include the critical window of sensitivity? Was the duration and frequency of exposure sensitive for detecting the endpoint of interest?	Considerations for this domain are highly variable depending on the endpoint(s)/outcome(s) of interest and must be refined by assessment teams. A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. *Good*: The duration and frequency of the exposure was sensitive and the exposure included the critical window of sensitivity (if known). *Adequate*: The duration and frequency of the exposure was sensitive and the exposure covered most of the critical window of sensitivity (if known). *Deficient*: The duration and/or frequency of the exposure is not sensitive and did not include the majority of the critical window of sensitivity (if known). These limitations are expected to bias the results towards the null. *Critically deficient*: The exposure design was not sensitive and is expected to strongly bias the results towards the null. The rationale should indicate the specific concern(s).

Questions used to guide the development of criteria for each domain in experimental animal toxicology studies

Evaluation type	Domain– core question	Prompting questions	Basic considerations
Outcome measures and results display	**Endpoint sensitivity and specificity–** Are the procedures sensitive and specific for evaluating the endpoint(s)/outcome(s) of interest? *Note:* *Sample size alone is not a reason to conclude an individual study is critically deficient.*	For each endpoint/outcome or grouping of endpoints/outcomes in a study: Are there concerns regarding the specificity and validity of the protocols? Are there serious concerns regarding the sample size (see note)? Are there concerns regarding the timing of the endpoint assessment?	Considerations for this domain are highly variable depending on the endpoint(s)/outcome(s) of interest and must be refined by assessment teams. A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. Examples of potential concerns include: Selection of protocols that are insensitive or non-specific for the endpoint of interest. Use of unreliable methods to assess the outcome. Assessment of endpoints at inappropriate or insensitive ages, or without addressing known endpoint variation (e.g., due to circadian rhythms, estrous cyclicity, etc.). Decreased specificity or sensitivity of the response due to the timing of endpoint evaluation, as compared to exposure (e.g., short-acting depressant or irritant effects of chemicals; insensitivity due to prolonged period of non-exposure prior to testing).

Questions used to guide the development of criteria for each domain in experimental animal toxicology studies				
Evaluation type		Domain– core question	Prompting questions	Basic considerations
🟩	🟩	**Results Presentation–** Are the results presented in a way that makes the data usable and transparent?	For each endpoint/outcome or grouping of endpoints/outcomes in a study: Does the level of detail allow for an informed interpretation of the results? Are the data analyzed, compared, or presented in a way that is inappropriate or misleading?	Considerations for this domain are highly variable depending on the outcomes of interest and must be refined by assessment teams. A judgment and rationale for this domain should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. Examples of potential concerns include: Non-preferred presentation, such as developmental toxicity data averaged across pups in a treatment group, when litter responses are more appropriate. Failing to present quantitative results. Pooling data when responses are known or expected to differ substantially (e.g., across sexes or ages). Failing to report on or address overt toxicity when exposure levels are known or expected to be highly toxic. Lack of full presentation of the data (e.g., presentation of mean without variance data; concurrent control data are not presented).

Questions used to guide the development of criteria for each domain in experimental animal toxicology studies

Evaluation type	Domain– core question	Prompting questions	Basic considerations
Overall Confidence	**Overall Confidence–** Considering the identified strengths and limitations, what is the overall confidence rating for the endpoint(s)/outcome(s) of interest? *Note:* *Reviewers should mark studies that are rated lower than high confidence only due to low sensitivity (i.e., bias towards the null) for additional consideration during evidence synthesis. If the study is otherwise well-conducted and an effect is observed, the confidence may be increased.*	For each endpoint/outcome or grouping of endpoints/outcomes in a study: Were concerns (i.e., limitations or uncertainties) related to the reporting quality, risk of bias, or sensitivity identified? If yes, what is their expected impact on the overall interpretation of the reliability and validity of the study results, including (when possible) interpretations of impacts on the magnitude or direction of the reported effects?	The overall confidence rating considers the likely impact of the noted concerns (i.e., limitations or uncertainties) in reporting, bias and sensitivity on the results. A confidence rating and rationale should be given for each endpoint/outcome or group of endpoints/outcomes investigated in the study. *High confidence*: No notable concerns are identified (e.g., most or all domains rated Good). *Medium confidence*: Some concerns are identified, but expected to have minimal impact on the interpretation of the results. (e.g., most domains rated Adequate or Good; may include studies with Deficient ratings if concerns are not expected to strongly impact the magnitude or direction of the results). Any important concerns should be carried forward to evidence synthesis. *Low confidence*: Identified concerns are expected to significantly impact on the study results or their interpretation (e.g., generally, Deficient ratings for one or more domains). The concerns leading to this confidence judgment must be carried forward to evidence synthesis (see note). *Uninformative*: Serious flaw(s) that make the study results unusable for informing hazard identification (e.g., generally, Critically Deficient rating in any domain; many Deficient ratings). Uninformative studies are not considered further in the synthesis and integration of evidence.

[a] For non-targeted or screening-level histopathology outcomes often used in guideline studies, blinding during the initial evaluation of tissues is generally not recommended as masked evaluation can make "the task of separating treatment-related changes from normal variation more difficult" and "there is concern that masked review during the initial evaluation may result in missing subtle lesions." Generally, blinded evaluations are recommended for targeted secondary review of specific tissues or in instances when there is a pre-defined set of outcomes that is known or predicted to occur (Crissman et al., 2004).